

# Projekt z SQL - Průvodní listina

## k 7 SQL souborům obsahujícím řešení výzkumných úloh

kurz: Datová akademie 26.10.2023

Zuzana Ševčíková

---

### Primární tabulka:

**Tabulky pojmenujte `t_{jmeno}_{prijmeni}_project_SQL_primary_final` (pro data mezd a cen potravin za Českou republiku sjednocených na totožné porovnatelné období – společné roky) a `t_{jmeno}_{prijmeni}_project_SQL_secondary_final` (pro dodatečná data o dalších evropských státech).**

### Volba východiskových tabulek

Ze zadání Výzkumných otázek Projektové úlohy vyplývají informace, které mají být obsaženy ve výsledné tabulce / tabulkách:

- 1) Trend mezd v jednotlivých odvětvích v čase,
- 2) Vztah mezi cenami, mzdami a časem (tím pádem musí jít o čisté mzdy),
- 3) Ceny potravin v čase,
- 4) Vývoj cen a mezd (čistých) v čase,
- 5) Vliv změny HDP na změnu mezd (čistých) a cen potravin.

Z tabulek dostupných pro vytvoření pracovních tabulek, a následně pro řešení 5 projektových úloh vyplývá, že porovnávání cen potravin, mezd a HDP je z dostupných dat možné jenom pro ČR. Z toho plyne, že primární tabulka musí být podkladem pro vypracování všech 5 úkolů. Je nutné pracovat s tabulkami `czechia_payroll`, `czechia_payroll_unit` a `czechia_industry_branch` (mzdy v jednotlivých odvětvích v čase), a spojit je s tabulkami `czechia_price`, `czechia_region` a `czechia_price_category` (obsahují informace o cenách vybraných potravin v čase v jednotlivých krajích ČR). K výše uvedeným tabulkám je potřeba připojit i tabulku `economies`, a ní vybrat informace týkající se některých ekonomických ukazatelů ČR v čase. Důvod je blíže vysvětlen v další kapitole.

## Výběr informací z východiskových tabulek do výsledné tabulky

Z tabulky czechia\_payroll nás zajímají jenom informace o „Průměrné mzdě na zaměstnance“. Tyto informace jsou napříč tabulkou definovány pod kódem 5958 ve sloupci „value\_type\_code“.

Tabulka czechia\_payroll pak obsahuje data mezd pro dva calculation\_code (100 a 200). Nepodařilo se mi získat zdrojová data s popisem významu sloupců, ani nemám znalosti k interpretaci těchto kódů. Podařilo se mi ale získat informaci (zdroj [www.cszo.cz](http://www.cszo.cz)), která mluví o průměrné **hrubé měsíční nominální mzdě** v národním hospodářství v roce 2020 ve výši 35402 Kč. Když v tabulce czechia\_payroll zprůměruji mzdy napříč odvětvími pro rok 2020 přes oba kódy (100 a 200), získám hodnotu 35741 Kč dotazem:

```
SELECT avg(value)
FROM czechia_payroll cp
WHERE value_type_code = 5958 AND industry_branch_code IS NOT NULL
      AND payroll_year = 2020;
```

Odchylka od publikované částky je přijatelně malá, a proto jsem se rozhodla dále pracovat se mzdami průměrovanými v jednotlivých letech přes všechny kvartály a oba kalkulační kódy:

```
SELECT payroll_year, industry_branch_code , avg(value),
       (sum(value) / count(payroll_quarter)), count(1)
FROM czechia_payroll cp
WHERE value_type_code = 5958 AND industry_branch_code IS NOT NULL
      AND payroll_year = 2020
GROUP BY payroll_year, industry_branch_code ;
```

jako s **hrubými mzdami v jednotlivých odvětvích v jednotlivých letech**. Čisté mzdy pak získám využitím informace o výšce daní v jednotlivých letech pro ČR z tabulky economies.

## Způsob spojování východiskových tabulek

V tabulce czechia\_payroll jsou data s kódem 5958 pro „Průměrné mzdě na zaměstnance“ od roku 2000 do roku 2021. V tabulce czechia\_price jsou data od 2006 do 2018, a v tabulce economies jsou pro vybrané sloupce data k dispozici spojitě od roku 2004 do 2018. Jelikož je informace o čase jediným pojivem všech zmíněných tabulek, výsledná tabulka je poskládána přes sloupce s informací o čase. Obsahuje tedy data od roku 2006 do roku 2018.

Problém v tomto směru představuje tabulka czechia\_price, ve které časové sloupce date\_from a date\_to představují začáteční a konečné datумы jednotlivých sérií zjišťování cen daných potravin v jednotlivých regionech ČR. Těchto sérií je v jednotlivých rocích mnoho, a zbytečně „nafukují“ tabulku. Jestli z ní vyberu data pro „mléko“ v roce 2006, vrátí mě 750 záznamů nerovnoměrně rozdělených mezi jednotlivé kraje. Proto jsem ceny jednotlivých potravin získané mnoha sériemi zjišťování zprůměrovala pro jednotlivé regiony v daném roku:

```
SELECT YEAR(date_from) AS `year`, category_code, region_code,
       round((avg(value)), 2) AS avg_price
FROM czechia_price cp
GROUP BY YEAR(date_from), category_code, region_code ;
```

Toto pak následně pro „mléko“ a rok 2006 vrátí 342 řádků.

Vybrané informace z tabulek `czechia_price` s číselníky a tabulku `economies` spojím do pohledu *v\_druha*. Vybrané informace z tabulky `czechia_payroll` s číselníky spojím do CTE s názvem *prva*. Obě skupiny tabulek následně spojím pomocí LEFT JOIN znovu na základě času do výsledné tabulky *t\_zuzana\_sevcikova\_project\_sql\_primary\_final*.

## Sekundární tabulka:

**Tabulky pojmenujte ... `t_{jmeno}_{prijmeni}_project_SQL_secondary_final` (pro dodatečná data o dalších evropských státech).**

Jelikož zdrojem dat pro vypravování 5 úkolů je primární tabulka, účelem té sekundární je už jenom poskládání dodatečných dat pro další evropské státy – tedy splnění výše uvedeného zadání.

Údaje o státech mimo ČR jsou v tabulkách `countries` a `economies`. Tabulka `countries` je, soudě z dat (sloupec s názvem `median_age_2018`) jenom pro rok 2018, a nelze z ní pak použít data měnící se v čase (a ani neobsahuje data použitelná ve výzkumných úlohách). Tabulka `economies` obsahuje data pro ČR, které jsou vybrané do `primary_table`. Data z tabulky `economies` pro ostatní evropské státy nejsou použitelná v projektových úlohách, protože neobsahují informace o cenách potravin a mzdách použitelné pro porovnání.

Pod termínem „Evropské státy“ neuvažuji o extrateritoriálních a zámořských státech a koloniích evropských států.

## 1. výzkumná otázka

**Rostou v průběhu let mzdy ve všech odvětvích, nebo v některých klesají?**

V tabulce `t_zuzana_sevcikova_project_sql_primary_final` pracujeme se sloupci `payroll_year`, `industry_branch` a `year_averaged_payroll_czk`, abychom zjistili trend vývoje mezd v průběhu let 2006-2018 ve všech 19 průmyslových odvětvích.

Při tvorbě tabulky `t_zuzana_sevcikova_project_sql_primary_final` spojováním se zmnožili řádky původní tabulky `czechia_payroll`, proto SELECT DISTINCT.

Potřebuji data seřadit podle roku a odvětví, a zároveň ve výstupu mít kromě těchto dvou sloupců ještě jeden, který by vypovídal o trendu v mzdách. Abych mohla použít GROUP BY podle roku a odvětví, tak ten třetí sloupec v SELECTu musí být nějak agregován. Nepotřebuji ale klasickou agregační funkci, potřebuji jenom jiné, přehlednější (pro účely porovnávání), vyjádření nárůstu mzdy. Ve výsledku je to vyřešené přes výpočet relativního nárůstu mzdy,

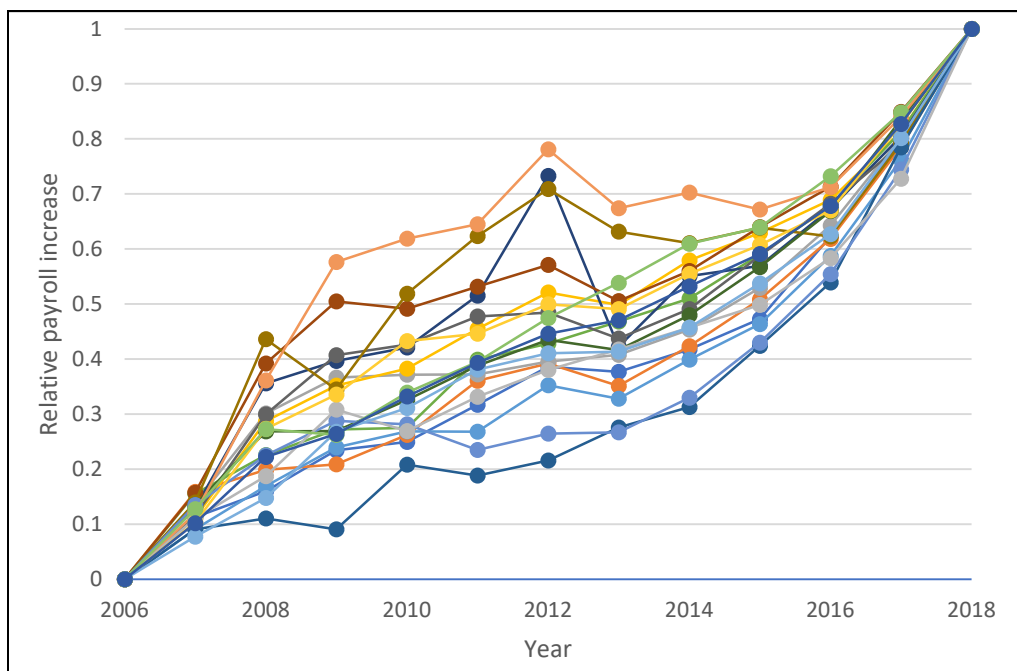
$$(\text{mzda v aktuálním roce} - \text{mzda v roce 2006}) / (\text{mzda v roce 2018} - \text{mzda v roce 2006})$$

který je v dotazu rozepsaný do posuzování hodnoty čitatele vůči jmenovateli jako kritérium při tvorbě nového sloupce přes CASE.

Výsledek:

V průběhu let ve všech sledovaných odvětvích mzdy v globálu rostly. Pro názornost viz obrázek níže, kde je vynesena závislost hodnoty relativního nárůstu mzdy v čase. Jde o bezrozměrné relativní číslo, nezohledňuje tedy absolutní výši mezd v daném odvětví. Tyto data vypovídají jen o tom, v kterém časovém období se mzda v daném odvětví zvedala víc, ve kterém míň, ve kterém dokonce dočasně klesala.

Tento posledně zmiňovaný trend je zajímavý třeba v sektoru Peněžnictví a pojišťovnictví, nebo ve Výrobě a rozvodu elektřiny, plynu, tepla, nebo Těžba a dobývání, kde do roku 2012 mzdy rostly relativně výrazně, a pak nastala stagnace, případně dočasný pokles. Původní tabulka czechia\_payroll neměla vazbu na regiony ČR, o této vazbě data nevypovídají.



## 2. výzkumná otázka

**Kolik je možné si koupit litrů mléka a kilogramů chleba za první a poslední srovnatelné období v dostupných datech cen a mezd?**

Předpoklady:

- první a poslední srovnatelné období = rok 2006 a rok 2018 (důvod vysvětlen v kpt. Způsob spojování východiskových tabulek).
- v tabulce t\_zuzana\_sevcikova\_project\_sql\_primary\_final se pracuje s hrubými mzdami v jednotlivých odvětvích v jednotlivých letech. Čisté mzdy získám využitím informace o výšce daní v jednotlivých letech pro ČR (info: viz kpt. Výběr informací z východiskových tabulek do výsledné tabulky).
- všechny životní náklady se mohou promítnout výlučně a jen do množství chleba / mléka, které si zaměstnanec v daném odvětví a daném regionu ČR může za svoji mzdu koupit.

Řešení jsem pojala jako hledání takového odvětví, kde je v roce 2006, resp. v roce 2018 nejnižší, resp. nejvyšší mzda v rámci ČR, a tudíž potenciál koupit si nejnižší, resp. nejvyšší množství potravin, které se ale zároveň mezi regiony ČR liší cenou. Všechny ostatní odvětví / regiony jsou pak někde mezi těmito extrémy.

V prvním kroku se vytvoří view `v_zs_task2`, kde si připravím sloupce, se kterými budu dál pracovat, včetně `net_wages` (čisté mzdy) a `amount_to_buy`, a zvolím rok a potravinovou kategorii. Toto mi umožní zmenšit velikost původní tabulky `t_zuzana_sevcikova_project_sql_primary_final`, protože zpracování druhého kroku (druhého dotazu), i takto trvá kolem 8-10 minut.

Ve druhém kroku hledám regiony a odvětví, kterým přináležejí maximum a minimum mezd a nakoupených chlebů / mlék. Zvolila jsem přístup přes `subselect` s `unionem`, protože jakákoliv varianta dotazu s `GROUP BY` mi tady vracela nesmysly.

Výsledky:

- A) Kde jsou v roce 2006 extrémy ve mzdách, a kolik nejvíc a nejméně litrů mléka si za tyto mzdy lze koupit?

region_name	industry_branch	net_wages [CZK]	amount_to_buy [L]
Jihočeský kraj	Ubytování, stravování a pohostinství	9770.1	636.1
Moravskoslezský kraj	Peněžnictví a pojišťovnictví	34045.3	2507.0

- B) Kde jsou v roce 2006 extrémy ve mzdách, a kolik nejvíc a nejméně kg chleba si za tyto mzdy lze koupit?

region_name	industry_branch	net_wages [CZK]	amount_to_buy [kg]
Karlovarský kraj	Peněžnictví a pojišťovnictví	34045.3	2232.5
Moravskoslezský kraj	Ubytování, stravování a pohostinství	9770.1	549.5

- C) Kde jsou v roce 2018 extrémy ve mzdách, a kolik nejvíc a nejméně litrů mléka si za tyto mzdy lze koupit?

region_name	industry_branch	net_wages [CZK]	amount_to_buy [kg]
Karlovarský kraj	Informační a komunikační činnosti	47777.8	2628.0
Pardubický kraj	Ubytování, stravování a pohostinství	15985.1	717.8

- D) Kde jsou v roce 2018 extrémy ve mzdách, a kolik nejvíc a nejméně kg chleba si za tyto mzdy lze koupit?

region_name	industry_branch	net_wages [CZK]	amount_to_buy [L]
Karlovarský kraj	Informační a komunikační činnosti	47777.8	2088.2
Olomoucký kraj	Ubytování, stravování a pohostinství	15985.1	611.1

Původní tabulka czechia\_payroll neměla vazbu na regiony ČR, proto informace z této tabulky se do výsledku promítne jenom přes vyselektování odvětví, ve kterém jsou nejvyšší nebo nejnižší mzdy v daném roce. Vazba na selekci regionu, ve kterém je možné si nakoupit za tyto mzdy více nebo méně vybraných potravin plyne z původní tabulky czechia\_price přes různé ceny těchto komodit v různých regionech. Proto se kraje, ve kterých lze nakoupit nejvíc / nejmíň potravin, ve výsledcích můžou lišit, i když mzdy ne.

### 3. výzkumná otázka

**Která kategorie potravin zdražuje nejpomaleji (je u ní nejnižší procentuální meziroční nárůst)?**

Předpoklady:

- region\_code (resp. region\_name) IS NULL pro celorepublikový průměr cen potravinových kategorií v původní tabulce czechia\_price,
- tabulka t\_zuzana\_sevcikova\_project\_sql\_primary\_final obsahuje data od 2006 do 2018 (důvod vysvětlen v kpt. Způsob spojování východiskových tabulek), tj. zpracovávám relativně nejnovější data, která v ní jsou: roky 2015-2018.

Řešení jsem pojala jako výpočet aritmetického průměru několika meziročních procentuálních nárůstů cen jednotlivých komodit. Z meziročních procentuálních nárůstů cen za roky 2015-2018 už by mohlo být vidět, jestli nejde jenom jakýsi sezónní úlet, ale o setrvalejší tendenci. Aritmetickým průměrem těchto procentuálních nárůstů pro jednotlivé potravinové kategorie jsem hledala tu, která dle zadání „zdražuje nejpomaleji (je u ní nejnižší procentuální meziroční nárůst)“. To chápu jako nejnižší kladný průměr uvedených tří meziročních procentuálních cenových nárůstů (viz vyznačený řádek v následující tabulce).

Výsledek:

food_category	avg_annual_price_increase_percent
Máslo	10.83
Vejce slepičí čerstvá	8.16
Eidamská cihla	7.12
Jablka konzumní	6.7
Mrkev	6.48
Konzumní brambory	4.47
Pomeranče	4.07
Šunkový salám	4.05
Pečivo pšeničné bílé	3.73
Pivo výčepní, světlé, lahvové	3.43
Hovězí maso zadní bez kosti	3.23
Rostlinný roztíratelný tuk	2.81
Chléb konzumní kmínový	2.76
Jakostní víno bílé	2.7
Jogurt bílý netučný	2.62
Těstoviny vaječné	2.54

Vepřová pečeně s kostí	2.53
Kapr živý	2.42
Rajská jablka červená kulatá	1.52
Mléko polotučné pasterované	0.73
<b>Rýže loupaná dlouhozrná</b>	<b>0.24</b>
Kuřata kuchařská celá	-0.12
Přírodní minerální voda uhlíčitá	-0.18
Papriky	-1.6
Banány žluté	-2.76
Pšeničná mouka hladká	-2.89
Cukr krystalový	-2.93

Řešením je SQL dotaz, který má dvě hlavní části. Prvá je CTE `cte_zs_task3`, ve které zužuji rozsah dat tabulky `t_zuzana_sevcikova_project_sql_primary_final`. Zejména snižuji počet sloupců na 3 a odstraňuji v nich znásobené řádky. Všechno za podmínky, že `region_name` is NULL, protože to by měla být data reprezentující celorepublikový průměr cen potravinových kategorií. Druhá část dotazu je výpočetní. Tady nejdříve v subselectu pospojím vícero tabulek z `cte_zs_task3` dokopy, mimo jiné za podmínky:

```
cte1.`year` = cteN.`year` + (N-1) AND cte1.`year` = 2018 , kde N je 2,3,4.
```

kterou se na první tabulku další připojuje za podmínky posunu roku o (N-1) níž. Vypočtu si meziroční procentuální nárůsty cen jednotlivých komodit. Z meziročních procentuálních nárůstů cen za roky 2015-2018 pak v selectu nad tímto subselectem vypočtu aritmetický průměr těchto meziročních procentuálních nárůstů cen jednotlivých komodit po rozřídění (GROUP BY) podle `food_category`.

## 4. výzkumná otázka

**Existuje rok, ve kterém byl meziroční nárůst cen potravin výrazně vyšší než růst mezd (větší než 10 %)?**

Předpoklady:

- `region_code` (resp. `region_name`) IS NULL pro celorepublikový průměr cen potravinových kategorií v původní tabulce `czechia_price`, tabulka `czechia_payroll` je zpracovaná jako celorepublikový průměr mezd,
- tabulka `t_zuzana_sevcikova_project_sql_primary_final` obsahuje data od 2006 do 2018 (důvod vysvětlen v kpt. Způsob spojování východiskových tabulek), tj. zpracovávám toto časové období.

Tuto úlohu jsem řešila paralelně dvěma způsoby – řešení **verze1** je použitelné jako podpůrný zdroj detailnějších informací k hlavnímu řešení – **verze2**:

Řešení **verze1** poskytuje porovnání meziročních nárůstů cen jednotlivých potravinových kategorií s meziročními nárůsty mezd v jednotlivých odvětvích od 2006, resp 2007 (protože meziroční ukazatel je k dispozici až ve druhém zahrnutém roce) do 2018, a vrací ty potraviny v těch letech a odvětvích, kde je splněna podmínka vyššího než 10-procentního rozdílu meziročních ukazatelů. Tím je tato

tabulka poněkud nepřehledná, ale po dalším dotazování je zdrojem podrobných informací ve spojení s hlavním řešením (verze2) tohoto úkolu.

**Hlavní řešení (verze2)** poskytuje porovnání meziročního nárůstu cen, získaného z cen, které jsou průměrem všech potravinových komodit a všech regionů, s meziročním procentuálním nárůstem mezd, které jsou celorepublikovým průměrem mezd ze všech odvětví. Pro roky od 2006 (resp. 2007) do 2018 vrací informaci, jestli je pro jednotlivé roky splněna podmínka vyššího než 10-procentního rozdílu meziročních ukazatelů.

Výsledek řešení verze2:

year	annual_price_increase_percent	annual_payroll_increase_percent	is_higher_10percent
2006	NULL	NULL	0
2007	6.76	6.84	0
2008	6.18	7.87	0
2009	-6.41	3.16	0
2010	1.95	1.95	0
2011	3.35	2.3	0
2012	6.73	3.03	0
2013	5.1	-1.56	0
2014	0.74	2.56	0
2015	-0.55	2.51	0
2016	-1.19	3.65	0
2017	9.63	6.28	0
2018	2.17	7.62	0

Z tabulky, která je hlavním řešením verze2 úkolu, je zřejmé, že neexistuje rok v rámci sledovaného časového rozpětí, ve kterém by byla splněna podmínka vyššího než 10-procentního rozdílu meziročních ukazatelů cen potravin a mezd v rámci ČR.

#### Popis dotazu hlavního řešení (verze 2):

SQL dotaz má dvě linie CTE, ve kterých jsou předpřipraveny tabulky na následné spojení pomocí vnitřního JOINu.

Prvá linie – CTE cte\_zs\_task43 – předpřipravuje data cen potravin. Nejprve z tabulky t\_zuzana\_sevcikova\_project\_sql\_primary\_final vyfiltruje jenom ty sloupce, se kterými se dál pracuje a odstraní znásobené řádky, aby v nadřazeném SELECTu fungovalo seskupování dat podle roku. V tomto nadřazeném SELECTu se vypočítá pro každý rok cenový průměr celé skupiny potravin. Funkce count(1) je tady jenom pro kontrolu fungování seskupení dat. V dalším nadřazeném SELECTu je zadefinován cenový průměr celé skupiny potravin pomocí funkce LAG a vypočítán meziroční nárůst cen potravin.

Druhá linie – CTE cte\_zs\_task44 – předpřipravuje data mezd obdobně jako je popsáno výše pro data cen.

Obě CTE linie se spojí vnitřním JOINem za podmínky rovnosti roku v cenové a mzdové části dotazu a přidá se nový sloupec, který zkoumá splnění podmínky vyššího než 10-procentního rozdílu meziročních ukazatelů cen potravin a mezd.







































## 5. výzkumná otázka

**Má výška HDP vliv na změny ve mzdách a cenách potravin? Neboli, pokud HDP vzroste výrazněji v jednom roce, projeví se to na cenách potravin či mzdách ve stejném nebo následujícím roce výraznějším růstem?**

Předpoklady: stejné, jako ve 4. výzkumné otázce.

Vnitřní logika dotazu je stejná jako ve 4. výzkumné otázce, verze2, protože jde o rozšíření dotazu 4. výzkumné otázky o třetí pomocnou CTE – cte\_zs\_task55 – která předpřipravuje data ohledně HDP, a tyto jsou pak vnitřním JOINem připojena na CTE cen potravin a mezd.

Výsledek:

year	annual_GDP_increase percent	annual_price_increase percent	annual_payroll_increase percent
2006	NULL	NULL	NULL
2007	 5.57	 6.76	 6.84
2008	 2.69	 6.18	 7.87
2009	 -4.66	 -6.41	 3.16
2010	 2.43	 1.95	 1.95
2011	 1.76	 3.35	 2.3
2012	 -0.79	 6.73	 3.03
2013	 -0.05	 5.1	 -1.56
2014	 2.26	 0.74	 2.56
2015	 5.39	 -0.55	 2.51
2016	 2.54	 -1.19	 3.65
2017	 5.17	 9.63	 6.28
2018	 3.2	 2.17	 7.62

Podívám-li se laickým okem na tuto tabulku ve smyslu otázky „má-li výraznější změna HDP odezvu v cenách potravin a v mzdách“, je zjevné, že propojení je přímější mezi HDP a mzdami, než mezi HDP a cenami potravin (případně mezi mzdami a cenami potravin).

Meziroční procentuální změna HDP přes 3% patrně vyvolává změny v meziročním platovém ukazateli. Zdá se, že ve sloupci ukazatele HDP rokem 2008 končí jakési „lepší období“, a začíná stagnace, která končí až roky 2014/2015. Opětovný meziroční růst HDP pak přichází rokem 2015, a pokračuje až do 2018.

Tento trend je víceméně kopírován ukazatelem meziročních změn mezd s ročním zpožděním.

Porovnám-li grafické trendy v HDP a mzdách, a naproti tomu trendy v HDP a cenách potravin, tak do druhého vztahu zřejmě vstupují další proměnné, protože propojení HDP – ceny potravin, případně mzdy – ceny potravin už z těchto dat tak zjevné není.